UMR 1041 INRA – AGROSUP

CESAER

Centre d'Economie et Sociologie
appliquées à l'Agriculture et aux Espaces Ruraux

Empirical validity of the evaluation of public policies:

models of evaluation and quality of evidence

Marielle Berriet-Solliec

Pierre Labarthe

Catherine Laurent

Jacques Baudry

Working Paper

2011/4

UMR 1041 CESAER INRA – AGROSUP
26, Bd Dr Petitjean - BP 87999 - 21079 Dijon cedex

Paper prepared for the 122nd EAAE Seminar
"EVIDENCE-BASED AGRICULTURAL AND RURAL POLICY MAKING: METHODOLOGICAL AND EMPIRICAL CHALLENGES OF POLICY EVALUATION"
Ancona, February 17-18, 2011

Empirical validity of the evaluation of public policies: models of evaluation and quality of evidence

Marielle Berriet-Solliec [1], Pierre Labarthe [2], Catherine Laurent [3] and Jacques Baudry [4]

1 Agro-Sup / INRA, Umr 1041, Dijon, France
2,3 INRA /Agroparitech, Umr 1048, Paris, France
4 INRA, Rennes, France

Pierre.Labarthe[at]agroparistech.fr

Empirical validity of the evaluation of public policies: models of evaluation and quality of evidence

Marielle Berriet-Solliec[1], Pierre Labarthe" Catherine Laurent[1] and Jacques Baudry

Abstract

There is a wide range of evaluation methods. On the basis of which criteria should one method be chosen? On what scientific foundations are the results of numerous evaluations based? How can reliability and empirical validity be tested? The relevance of such questions is heightened in new fields of action such as agri-environmental policy. This paper aims to demonstrate that theoretical advances on level and types of evidence (existence, causality, effectiveness) can help resolve these issues. The main evaluation methods are classified into three main categories, according to their main goal (to learn, measure, understand) and linked to the debate on types of evidence. The analysis is illustrated by comparing evaluation methods in the field of agro-environmental policies and farm advisory services. Attention is drawn to the shortcomings of each method with respect to corroborating facts as well as existing complementarities and trade-offs between methods in terms of empirical validity.

Keywords: evaluation, evidence, agricultural extension, agri-environment

JEL classification:
B49 - Economic Methodology - Other
H83 - Public Administration; Public Sector Accounting and Audits
Q18 - Agricultural Policy; Food Policy
Q58 - Environmental economics - Government Policy

- INTRODUCTION

Evaluation is a critical component of efforts to better target public action. For public decision makers, however, the existence of a wide range of evaluation methods raises a number of questions. On the basis of which criteria should an evaluation method be chosen? On what scientific foundations are the results of numerous evaluations based? How can reliability and empirical validity be tested? The relevance of such questions is heightened in new fields of action such as agri-environmental policy, where knowledge gaps and a lack of statistical data exist (Laurent 2007).

This paper aims to demonstrate that theoretical advances on types and levels of evidence can help resolve these issues. We begin by reviewing the main approaches used in evaluative research in the last 40 years to examine the main issues concerning these approaches. This

review serves as the basis for a simple classification of different evaluation methods to be linked to the debate on types of evidence. Our analysis is illustrated by comparing evaluation methods in the field of agro-environmental policies and agricultural advisory services. Attention is drawn to the shortcomings of each method with respect to corroborating facts as well as existing complementarities and trade-offs between methods in terms of this empirical validity.

- 1. DIVERSITY OF EVALUATION APPROACHES

There is a wide range of literature on public action programme evaluation based on an extensive variety of methodological positions and models. Reference works such as Rossi and Freeman, 1979/2004 and Patton, 1978/2008, which are republished on a regular basis, present the various aspects of this literature and the numerous questions it raises. Such reviews are a reminder that studies on evaluation reflect two recurrent goals for final users of the results of an evaluation procedure: to assess the costs and benefits of public action programmes, and to confirm the relevance of the theories underlying evaluation approaches.

1.1. Range of questions examined

The first issue indicates a need to look at the evaluation process in terms of accountability. This involves measuring the value of goods or services produced through public action programmes and comparing this value against the cost of their production. The goal is to determine whether an organisation or initiative has produced as many benefits as possible with the goods it has acquired or produced given the resources it has at its disposal. This approach takes into account a combination of factors: costs, quality, use of resources, appropriateness and respect of deadlines.

However these tenets are general and can be expressed in many different ways. For instance they can be translated into the idea of "value for money" developed in the early 1980s and appropriated by the New Public Management stream. Cost-benefit analyses are one of the main tools used in this case. Result models which focus on results and performance, are others.

Many approaches exist – and this diversity may be confusing to users. This difficulty is exacerbated by the similarly wide range of theoretical models upon which the public action programmes under evaluation are developed and implemented (logic intervention, theory of action, theory of programme, theory driven evaluation ...). This phenomenon is so widespread that Patton (2008) refers to a state of "Babel confusion".

In any case, at different levels of analysis, the central goal of evaluation research is to organise and analyse information gathered on the public action programme under evaluation to identify patterns which can be used to analyse the main types of public action (Shadish *et al.*, 1991). Emphasis may nevertheless be placed on different aspects. In a stylised way, research can be divided into three broad groups.

i) The first focuses on elements which support the use of evaluation. It seeks to facilitate the use of adequate methods and the appropriation of evaluation findings by different types of users (Patton 1975). Evaluation is considered an operational approach intended to improve public action programmes and decisions. Emphasis is placed on its instrumental dimension (response to an institutional demand) and on the role played by evaluation approaches as an organisational learning process.

ii) The second group of research focuses on the quantification of programme outcomes using micro-economic techniques (Rossi and Freeman, 1988). Several micro-economic studies, in line with the work of Heckmann, fall under this category. A founding principle of this type of research is the identification of an experimental situation in which systematic reference to a counterfactual can be used to identify outcomes which are specific to the programme under evaluation (Campbell *et al*., 1999).

iii) A third group is centred on the study of the theories underlying the public programme and analysing the specific mechanisms by which these programmes have produced results (Shadish *et al*. 1991, Chen, 1990). By resorting to theory, ex ante analytical frameworks can be created to structure, organise and prioritise evaluation results. Rather than a slip into abstraction, this approach aims to better understand the complexity of the subject under evaluation using a theoretical detour (Scriven, 1998). The development of critical evaluation (Pawson and Tilley, 1997) can be linked to this approach.

1.2. Typology of evaluation models

Scriven (1998), Pawson and Tilley (1997) focus on different aspects of evaluation because evaluation procedures themselves may have differing goals which must be kept in mind when assessing their quality and the pertinence and reliability of results. We have grouped these diverse evaluation methods into three types according to the goals assigned to the evaluation.

- [Goal 1: To learn]: the evaluation is primarily designed as a collective learning process;
- [Goal 2: To measure]: the evaluation is designed to assess programme performance and impact;
- [Goal 3: To understand]: the evaluation identifies and analyses the mechanisms by which the programme under evaluation can produce the expected outcomes or may create adverse effects.

Choosing an evaluation method in relation to these goals is essential. This can be done through a thoughtful examination of the recent debates on types and levels of evidence.

- 2. TYPES AND LEVELS OF EVIDENCE AND EVALUATION

When an evaluation procedure is used to assess a public action programme, generally the goal is to produce the best iknowledge possible on the actual impact of the programme. The ‘best’ knowledge should be a) based on relevant empirical evidence (which addresses the

question at hand); b) corroborated by facts (facts which reflect real world observation and are empirical in nature) and c) reliable (knowledge produced using rigorous methods).

2.1. Types of evidence

Defining the notion of empirical evidence is in itself a subject of debate (Cartwright 2007). Certain principles identified in the sphere of evidence-based medicine and its development, however, are useful in understanding the requirements that result from a careful corroboration of facts in an evaluation.

Broadly speaking, four types of empirical evidence may be necessary to evaluate public policies (Laurent *et al*. 2009):

i) Evidence of existence: the description and verification of facts which exist on the ground (e.g. a botanical inventory used to assess biodiversity conservation); this type of evidence is used to build an agreement among different actors on the state of the world (before and after the programme);

ii) Evidence of causality: produced when an event is necessary for an outcome (e.g. an increase in fertilizer, in a controlled environment, which increases crop yield). This type of evidence confirms a relationship of cause and effect between two variables, all other things being equal. It can be used to predict what mechanisms are needed for the goals of a public programme to be achieved;

iii) Evidence of effectiveness: evidence that a given action yields the desired result (e.g. improved indicators for biodiversity conservation following the implementation of agricultural regulations aimed at conservation). This type of evidence is needed to identify outcomes which are specific to a public programme: its actual impact on specified goals;

iv) Evidence of harmlessness: obtained when any negative effects of an action have been looked for and not found (e.g. a pesticide's effects on human health). This type of evidence attests to an absence of adverse effects a programme could have in other aspects not related to the programme's goal.

2.2. Levels of evidence

It should be noted, however, for each type of evidence – evidence of effectiveness, for example – not all findings can be ranked at the same "level of evidence", nor are they all equally reliable. Evidence can thus be prioritised according to its level of empirical validity. In the field of agricultural and sustainable development research, for example, the reliability of evidence can be classified in the following order, from least to greatest:

1. the opinions of respected authorities;
2. evidence obtained from historical or geographical comparisons;
3. evidence obtained from cohort studies or controlled case studies;
4. evidence obtained from gathering data on representative situations for hypothesis testing and statistical validation of the robustness of the results;
5. evidence obtained through randomised controlled trials (RCT).

The apparent simplicity of this classification should not conceal the numerous questions that arise, however, when several types of evidence are involved and need to be combined (Laurent, Trouvé 2011).

Ideally, an evaluation procedure should aim at producing results based on a high level of evidence for the types of evidence that are relevant to the procedure's goals. As demonstrated below, however, 'relevant' evidence varies depending on the purpose of the evaluation.

- 3. TYPE OF EVIDENCE FOR EVALUATIONS PRIMARILY DESIGNED AS A COLLECTIVE LEARNING PROCESS [TO LEARN]

3.1. General principles

The primary goal of certain evaluation procedures is to promote learning through close collaboration between different stakeholders right from the evaluation design stage. Their involvement in the evaluation is intended to build awareness and encourage new practices. The latter take precedence over the measurement of a programme's outcomes.

In line with this type of model are participative evaluation methods which highlight the educational dimension of evaluation procedures. These methods "bring to the table" all stakeholders who have a vested interest in the improvement of the programme under evaluation. The researcher begins by drafting as accurately as possible a sociogram of the network of stakeholders which includes information about the nature and intensity of the ties between these actors. This type of approach has a significant psychosociological dimension. The evaluator uses this representation of actor networks to conduct in-depth interviews with stakeholders to gather each person's point of view and suggest ways of improving the programme. At each stage of the evaluation, partial conclusions are discussed and analysed in working groups. In certain cases – service-related programmes, for example – evaluators constitute a representative sample of service users. It is the users themselves who then assess the value of the programme (after a specific training session).

Here, it is the contributions of programme stakeholders to a social construction of representations of an observed reality which is at the heart of the evaluation approach. While the process may simply mobilise opinions, it also calls upon scientific knowledge (in the field of natural sciences, primarily), often through the tools proposed by researchers (simulation models, etc.). The reliability of evidence used for collective learning is not frequently addressed in these approaches. Certain authors, however (Van der Sluijs *et al*. 2008) believe this process matters and influences the nature and content of the related learning.

In this type of evaluation procedure the evaluator's role is to organise debates, ultimately to obtain the most concordant results possible which can be used by the greatest number of people. These approaches have become highly popular in recent years and take on different forms. They are used for numerous issues involving collective action (water management, land-use planning) and rely on different methods to promote interaction among actors during the evaluation phase (role playing, multiagent-based simulation, etc.).

3.2. Implementation: the Soft System Methodology (SSM) example

An emblematic example of this type of approach can be found in evaluations of public programmes of farm advisory services In the last few years this process has been the subject of debate and growing importance has been attributed to participative conception and evaluation approaches. A notable example is the relatively widespread use of Soft System Methodology (SSM) to design and evaluate technical advisory programmes (Rohs and Navarro 2008). This method was used to evaluate public training programmes for advisory services in an agri-food supply chain in Thailand (Navarro *et al.* 2008). SSM is designed to help *human activity systems* (HAS) make the most effective decisions in uncertain and complex contexts (Checkland 1981): learning is the priority. Checkland and Scholes (1990) point out that SSM as a model is not intended to establish versions of reality. Instead, it aims to facilitate debate so that collective decisions and action can be taken in problem situations. The seven stages of SSM are (Checkland 1981): i) inquiring into situation (identifying the problem using different communication techniques: brainstorming, interviews, participant observation, focus groups, etc.); ii) describing the situation (describing the context using a wide variety of sources); iii) defining Human Activity Systems (HAS) (identifying programme stakeholders, and interviewing them on the transformations they are expecting); iv) building conceptual models of HAS (representing the relationships between stakeholders in the programme being designed or evaluated); v) comparing the conceptual models with the real world (preparation of a presentation of the model for a debate with stakeholders); vi) defining desirable and feasible changes; vii) implementation (Navarro *et al*. 2008).

Corroboration with facts and producing evidence do not appear to be at the heart of this conception/evaluation approach, which instead aims at promoting and structuring debate between programme stakeholders to arrive at a consensual solution. In practice, however significant problems arise (Salner 2000, Gonzalez and Benitez 2008). In workshops, for example, evidence is provided by different stakeholders verbally, and must be verified. Salner (2000) likens this method to journalism, in that it involves the verification of the opinions of different stakeholders so that "analysis makes it possible to mount an argument for change which was not simply an intuitive reaction to a conversation held; it was an argument which could be explicitly retraced at any time with links to supporting evidence" (Checkland and Sholes, pp. 198-99, cited by Salner 2000). Verification is thought to be guaranteed by the open, public and collective nature of the debate. Comparison with 'fact checking' in journalism, however, only holds true if the evidence presented is evidence of existence describing facts known through stakeholder practices. Instead, arguments often go deeper and target the expected or measured impact of programmes and even the causality pattern upon which they are based. In other words, these evaluation methods rely on evidence of effectiveness and causality but do not formalise this integration. This lack of formalisation manifests itself on two levels: (i) in the use of scientific knowledge to formulate hypotheses on the modalities of how public programmes function, (ii) in the verification of the level of evidence obtained.

Ultimately, these formalisation tasks are implicitly transferred to workshop leaders (often researchers). This situation poses a number of problems as it is assumed that these leaders have extensive skills and means at their disposal (to produce state-of-the art of scientific litterature, statistical analyses and various types of verifications). For this reason, several authors have pointed out that SSM may be exploited to reinforce a balance of power given the discrepancies in information between stakeholders: "the kind of open, participative debate that is essential for the success of the soft system approach, and is the only justification for the result obtained, is impossible to obtain in problem situations where there is a fundamental conflict between interest groups that have access to unequal power resources. Soft system thinking either has to walk away from these problem situations, or it has to fly in the face of its own philosophical principles and acquiesce in proposed changes emerging from limited debates characterized by distorted communication" (Jackson 1991, p. 198)

3.3. The evidence issue

These approaches raise several questions where the issue of evidence is concerned.

i) the issue of level of evidence is often neglected and seen as secondary to collective learning objectives. So that no participant is excluded, all contributions are viewed equally and the reliability of evidence is not subject to systematic testing procedures;

ii) very quickly, evidence presented by participants with different interests can be in competition and arbitration is often based on non-transparent criteria;

iii) without a systematic, clear verification procedure for evidence brought to the debate, learning may focus more on the ability to reach consensual positions than on the ability to use the best tools for achieving a given objective and on evaluating outcomes in a rational and rigorous manner.

- 4. TYPE OF EVIDENCE FOR EVALUATIONS DESIGNED TO MEASURE PERFORMANCE [TO MEASURE]

4.1. General principles

In evaluations that measure performance, the quantitative dimension takes precedence. Performance can be measured from many angles: economic (revenue, investment), technical (yields), or environmental (measuring biodiversity or pollution levels). This type of approach uses statistical tools to test the robustness of the driving forces of a programme (funding, investment, standards, etc.) on variables selected as proxy of a programme's goals (Bondonio 2005). The impact of a programme is defined as the difference between the actual situation with the programme and the situation without it.

The goal is to provide empirical evidence of the programme's effectiveness in order to measure as best as possible the actual impact of a public programme – and *only* that. In other words:

- the evaluation process is designed to measure final, actual performance and does not infer whether this performance is based on an underlying theory of action or not. For example, if an agri-environmental scheme is financed to protect biodiversity, performance evaluation will focus on the measure of biodiversity conservation, not on the underlying changes in farmers' practices;

- the evaluation process does not examine in detail the mechanisms by which an action is effective; public programmes mobilise a large number of factors and it is impossible to observe every form of interaction between them. Evidence of effectiveness is not sought in order to understand how the measure is effective. In this approach, the evaluator does not open the 'black box' of the evaluated programme. As a consequence, evidence that an agri-environmental scheme has been effective in maintaining biodiversity cannot be used to analyse the specific ecological, economic and social mechanisms that contributed to that outcome.

4.2. Implementation: measuring effectiveness independently of causality

Producing evidence of the effectiveness of a public action programme requires identifying that there is a relation "all other things are equal" between two variables: variable A, a proxy of the treatment or public programme under evaluation, and variable B, a proxy of the desired outcomes of that treatment on a population. Here, the main objective is to measure the difference between an observable situation (the level of the variable B, for the population that benefit from the treatment or programme, variable A) and a counterfactual unobservable one (the level of variable B which would still occur in this same population, but without the treatment, A). In practice, this is done by comparing the levels of variable B in a population which received the treatment and in a control population which did not. The main pitfall in this approach is a selection bias where differences exist between the 'treated' group and the control group (stemming from observable or unobservable factors) which could explain variations in levels of variable B independently of the effects of treatment A.

In light of this, EBP debate in the medical field proposes a hierarchy of the methods used in terms of their empirical reliability through their ability to reduce this bias. The smaller the bias, the higher the level of evidence. Traditionally, randomised control trials (RCT) are viewed as the 'gold method' for measuring the outcomes of a specific programme. Selection bias is eliminated by randomly distributing individuals in the treated group and the control group. While such methods are widespread in health-related fields, they are rarely used for other public programmes, where the randomisation of beneficiaries of a public programme can pose technical and ethical problems. For this reason, new, experimental evaluation approaches (Duflo and Krémer, 2005) are emerging in various sectors (justice, education, social sciences as well as

the environment and agriculture). In cases where RCT cannot be used, 'semi-experimental' methods such as matching[1] or double difference[2] are considered the most reliable alternatives.

Such methods have already been used to evaluate farm advisory service policies (Davis *et al.* 2010, Godtland *et al.* 2004). To ensure the empirical reliability of this kind of scientific work, methodological precautions must be taken which may limit the scope of findings. Below are two examples related to advisory service public programmes.

i) The first problem bears on the random distribution of farmers who benefited from these programmes of advisory services and those who did not (in the case of RCTs). Aside from the ethical issues raised, this requirement is also contrary to the diagrams of causality of certain programmes, such as participative and *bottom-up* approaches (e.g. farmers' field schools): the effectiveness of such programmes theoretically depends on the self-motivated participation of farmers in a collective project

ii) The second problem bears on an essential hypothesis of any evaluation of impact (RCT or semi-experimental evaluation): beneficiaries must not be influenced by the fact that non-beneficiaries do not benefit from the programme, and vice versa (Stable Unit Treatment Value Assumption – SUTVA). This hypothesis may also be contrary to the diagrams of causality underlying certain advisory service programmes, particularly those built on so-called diffusionist models (e.g. the World Bank's Train & Visit -T&V- programme). In theory, their effectiveness resides in the fact that farmers who directly receive advice can share acquired knowledge with those who have not.

Measurement of the impact of public programmes is rigorous if methods are used which are consistent with specific hypotheses (randomisation, a lack of diffusion-related effects, etc.). It is important to address the consequences of these hypotheses on the validity of an evaluation's findings, as well as how may complement other findings, for example in relation to the diagrams of causality of the policies under evaluation. In other words, the methods used to rigorously measure the impact of farm advisory policies are such that they cannot be used at the same time to determine whether the theoretical causality patterns of these programmes (e.g. knowledge sharing in T&V programmes) are empirically verified during the implementation phase.

1. [1] Matching involves pairing individuals who benefitted from the programme with individuals who did not and comparing the levels of indicator variables. The goal is to pair individuals based on their most significant similarity, particularly in terms of how likely they are to benefit from the programme.
2. [2] The double difference method is a combination of a comparison before and after the implementation of a public programme and a comparison with and without the programme. Researchers measure differing trends in variable B in both the beneficiary group and the control group. In environmental studies, this approach is known as BACI (before-after control impact) (Bro, 2004).

4.3. The evidence issue

Fully understanding the significance and limitations of these approaches is only possible if we accept that they are designed to obtain the highest possible level of empirical evidence of effectiveness (or harmlessness) – and only this.

i) obtaining evidence of effectiveness may seem simple or even simplistic. It is in fact quite challenging and involves costly practices which pose significant methodological and ethical problems. Nevertheless, it is the only way to obtain rigorous evidence of the actual impact of a public action programme;

ii) the significance of evidence of effectiveness obtained in this way should not be overestimated; it does not indicate precisely which mechanisms rendered programme action effective; it cannot demonstrate causality or corroborate theoretical statements; often, several competing explanations emerge concerning the effectiveness of a programme;

iii) these results are therefore of limited interest when deciding to extend a public action programme to other contexts or periods. This should be done using approaches which provide the most reliable hypotheses possible regarding the *mechanisms* that make action effective – in other words, on the basis of evidence of causality.

- 5. TYPE OF EVIDENCE FOR EVALUATIONS AIMED AT IDENTIFYING AND ANALYZING THE MECHANISMS BY WHICH A PROGRAMME CAN PRODUCE EXPECTED RESULTS OR ADVERSE EFFECTS [TO UNDERSTAND]

5.1. General principles

Several authors have pointed out the importance of using theory in evaluation approaches (Shadish *et al*, 1991). This has been stressed by authors from the school of thought on realistic evaluation in particular (Pawson and Tilley 1997). According to these authors, what is central in a realistic approach is a precise understanding of the mechanisms operating in the programmes being studied. Their initial acknowledgement underlines the limitations of the first two types of evaluation described above. They insist on the fact that these evaluations, which rely on gathering opinions, on role-playing games or on evidence of effectiveness, cannot reveal in a reliable way the causal relations that explain why a programme works or not in a given context.

Realistic evaluation focuses on understanding (i) the object which is evaluated (ii) the mechanisms of action to be 'revealed' through the analysis and (iii) the context in which the programme is implemented. By analyzing, in different contexts, the way in which the impacts are produced, regularities or recurring facts are identified so as to determine the causal relations by which the implementation of a public program has expected or unexpected effects. These effects can directly relate to the goal of the program or to its broader context. The evaluation will thus depend on the nature of the problem in question: what is at stake are the specificities of this problem in a given context and the assessment of the degree of genericity of the proposed solutions.

In certain cases, to improve the quality of the measurement of impacts, the model the evaluation is constructed using the preliminary analysis of the theory underlying the program (program theory). A first step is the understanding (before the measurement) of the causal mechanisms that guided the design of the program (Chen, 1990). The role of the evaluator consists, more precisely, in putting forth hypotheses on the causality patterns between a program and its potential subsequent effects. The aim is to build a diagram which traces these patterns of causality and constitutes the theory of the programme. When it is established, such a diagram becomes a reference framework and the basis of the evaluation approach for the evaluator, who proposes indicators that will be useful for measuring impacts. This way of using evidence of causality in the evaluation process raises many questions as we shall see in the following example.

5.2. The example of environmental evaluation and coupling between economic models and sustainability indicators

Many public programmes aim at encouraging farmers to adopt practices which guarantee better environmental performances (preservation of biodiversity, water quality, etc.). This is done by delivering specific financial support or by making changes in farm practices a prerequisite to receiving existing forms of aid (agri-environmental schemes, cross-compliance for the first pillar of the CAP).

The procedures used to evaluate the environmental impact of these programmes almost never rely on the production of evidence of effectiveness, as seen in Kleinj and Sutherland's review on biodiversity (2003). Measuring the effectiveness of a programme for biodiversity conservation would indeed require collecting ecological data according to a specific and elaborate methodological framework (with the possibility of building counterfactuals so as to measure impacts specifically linked to the programme). Such methodological frameworks are costly and often regarded as inaccessible. For this reason, many evaluations rely on the diagram of causality at the origin of the public programme (an economic incentive A, must cause a change of agricultural practice B, which has an ecological impact C), and make the assumption that if the means were in fact implemented, then the programme was effective. Evaluations then focus for instance on measuring the number of farmers who have actually changed their practices. This has led Primdahl *et al.* (2003) to speak about the measurement of "policy performances". In certain cases this information is considered sufficient to draw conclusions on the environmental impact of the programme. In other cases, the 'black box' of these changes is opened and additional data are collected (about crop rotation, plant pest management, etc). They are linked to agri-ecological indicators to calculate the potential risks and effects of these changes (for example the use of less chemical inputs is associated with a positive impact on the biodiversity) (Mitchell *et al.*1995, Van de Werf, Petit 2002).

The theoretical articles which examine these types of methods all underline the fact that these measures identify 'potential effects' but do not measure impacts. Nevertheless, these

precautions are often absent in the executive summaries of reports which present the results of these approaches. Variations in the value of an indicator can thus be presented as evidence of an improvement of environmental performances. This is not only improper from a formal point of view; the few experimental tests carried out on this issue also disprove that it is an acceptable estimate. For instance Kleinj and Sutherland (2003) and Kleinj *et al*. 2006 show that certain measures which were successful in terms of "policy performance", did not have the expected environmental impact.

Such doubts about the effectiveness of certain agri-environmental schemes can be linked to the weakness of the theoretical models upon which they are based, as well as to a lack of empirical data with which to identify what works and what does not (McNeely *et al*. 2005). This concerns both evidence of effectiveness and evidence of causality. This is why one may regret the limited development of realistic evaluations in the agri-environmental field. Such evaluations would be very useful in identifying the reasons why responses to a public programme vary according to the social and ecological contexts. The work done on the eco-millennium assessment showed a dramatic lack of such knowledge (Carpenter *et al*. 2006).

5.3. The evidence issue

Recourse to evidence of causality in evaluation procedures thus takes two principal forms: either one seeks to produce such evidence, and the emphasis is placed on its relevance to reality and its capacity to reveal in detail the mechanisms behind the phenomena observed; or one relies on evidence of causality to analyze the way in which it interacts in the theory of the program, in order to structure the evaluation consequently.

i) Both ways of using evidence of causality are fundamental: identifying the mechanisms by wich the actions were effective (or not) is essential to produce generic knowledge that can be used to develop new public programmes and also to raise new questions for the researchers and stakeholders involved in the evaluation.

ii) The use of theoretical models to infer the effective performance of a public programme, as sophisticated as they may be, is always limited; the causality patterns formalized in these theoretical models are always only partial representations of complex phenomena. Their predictive capacities are always limited and can vary according to the object under evaluation and the context; therefore one cannot replace the observation of the real effects (and the production of evidence of effectiveness) by that of expected effects (estimated using an analysis of the means implemented in the public programme).

- 6 CONCLUSION

Underlying the diversity of evaluation approaches are very different objectives: understanding the mechanisms of public programmes, measuring their specific impacts, or

supporting collective learning and the emergence of an agreement between stakeholders concerned by public action programmes. Each approach deals with different facts (mechanisms, impact, stakeholder participation); as such, the issue of fact corroboration and the reliability of the results of an evaluation process must be addressed in different terms.

A common feature of these approaches, however, is that they can only produce reliable results for a limited field of investigation. Clearly identifying the limited validity of the findings of each evaluation approach is thus a requirement to avoid erroneous interpretations. The analytical framework that we present in this paper to connect the goal of evaluations with types and levels of empirical evidence can facilitate this clarification and interpretation.

To be explicit about these issues has a double advantage. On the one hand it makes it possible to avoid certain errors by not using one type of evidence in place of another (for example by considering evidence of effectiveness as equivalent to evidence of causality). On the other hand, it opens new perspectives for a rational articulation of various types of evaluation approaches.

Ideally, one would like to be able to systematically combine the three objectives of the evaluation (to understand, measure, learn) in a rigorous overall framework. This ideal is often inaccessible, for reasons of cost, methodological impossibilities, etc. Furthermore it is often necessary to select in a drastic way very precise objectives for each evaluation from a large number of possible points of view.

Choices should thus be made. As is the case with any public policy instrument, these choices depend on a multiplicity of factors which cannot be reduced to a rational approach based on evidence (Weber 1921). But a better shared knowledge of the types of evidence that can be produced can help clarify what is at stake in making these choices.

ACKNOWLEDGMENTS

This paper is based on researches which were funded by the French National Research Agency (ANR) (Ebp-BioSoc and BipPop programs).

REFERENCES

Bondonio, D. (2002). Evaluating Decentralized Policies: A Method to Compare the Performance of Economic Development Program across Different Regions or States. *Evaluation*, 8(1): 101-124

Bro, E., Mayot, P., Corda, E. and Reitz, F. (2004), Impact of habitat management on grey partridge populations: assessing wildlife cover using a multisite BACI experiment. *Journal of Applied Ecology*, 41: 846–857.

Campbell, D., Russo, M.J., andJean M. (1999). *Social experimentation*, Thousand Oaks (CA) : Sage.

Carpenter, S., DeFries, R., Dietz, T., Mooney, H., Polasky, S., Reids, W., and Scholes, R. (2006).. Millenium Ecosystem assessment : research needs. *Science*, 314:257_258.

Cartwright, N. (with help from Goldfinch, A. and Howick, J.) (2007). Evidence-based policy: Where is our theory of evidence? Centre for Philosophy of Natural and Social Social Science, London School of Economics, Technical Report 07/07, 17 p.

Checkland, P.B., (ed.) (1981). *System thinking, system practice*. New-York : John Wiley.

Checkland, P.B., and Scholes, J. (eds.) (1990). *Soft systems methodology in action*. Chister (GB): John Wiley & sons.

Chen, H.T. (ed.) (1990). *Theory-Driven Evaluation*. Newbury, CA: Sage.

Davis, K., Nkonya, E., Kato, E., Mekonnen, D.A., Odendo, M., Miiro, R., and Nkuba, J. (2010). Impact of Farmer Field Schools on Agricultural Productivity and Poverty in East Africa IFPRI Discussion Paper 00992 June 2010 Knowledge, Capacity, and Innovation Division, IFPRI, Washington, 56 p.

Duflo, E., and Kremer, M. (2005). Use of Randomization in the Evaluation of Development Effectiveness. In Pitman, G., Feinstein, O., and Ingram, G. (eds.), *Evaluating Development Effectiveness*. New Brunswick (NJ): Transaction Publishers, 205-232.
European Commission (EC) (2009). Evaluation of the Implementation of the Farm Advisory System Final Report –

Gonzalez, C.R.N., and Benitez, S.O. (2008). Validación del intervenciones 'SOFT OR'. *Investigação Operacional*, 28(1): 59-75.

Godtland, E.M., Sadoulet, E., de Janvry, A., Murgai, R., and Ortiz, O. (2004). The Impact of Farmer Field Schools on Knowledge and Productivity: A Study of Potato Farmers in the Peruvian Andes. *Economic Development and Cultural Change*, 53(1): 63-92.

Jackson, M. (ed.) (1991). *Systems methodology for the management sciences*. New York and London: Plenum Press.

Kleinj, D., and Sutherland, W. (2003). W. How effective are European agri-environment schemes in conserving and promoting biodiversity? *Journal of Applied Ecology*, 40 : 947–969

Kleinj, D., Baquero, R. A., Clough, Y., Díaz, M., Esteban, J., Fernández, F., Gabriel, D., Herzog, F., Holzschuh, A., Jöhl, R., Knop, E., Kruess, A., Marshall, E. J. P., Steffan-Dewenter, I., Tscharntke, T., Verhulst, J., West, T. M., and Yela, J. L. (2006). Mixed biodiversity benefits of agri-environment schemes in five European countries. *Ecology Letters*, 9(3):243-254.

Laurent, L. (Ed.) (2007). Validité des connaissances scientifiques et intervention publique : le cas de l'agriculture dans le développement durable. Projet ANR « EBP-BIOSOC. 2007-2010» 17 p. + annexes. # 2007 ?

Laurent, C., and Trouvé, A. (2011). Competition of evidences and the emergence of the "evidence-based" or « evidence-aware » policies in agriculture. 122nd EAAE Seminar "evidence-based agricultural and rural policy making: methodological and empirical challenges of policy evaluation" Ancona, February 17-18, 2011

Laurent, C., Baudry, J., Berriet-Solliec, M., Kirsch, M., Perraud, D., Tinel, B., Trouvé, A., Allsopp, N., Bonnafous, P., Burel, F., Carneiro, M.-J ., Giraud, C., Labarthe, P., Matose, F., and Ricroch, A. (2009). Pourquoi s'intéresser à la notion d'Evidence-based policy ? *Revue Tiers-monde*, 200: 853-873.

Laurent, C., Cerf, M., and Labarthe, P. (2006). Agricultural extension and market regulation : learning form a comparison of six EU countries. *Journal of Agricultural Education and Extension*. 2006, 12(1) : 5-16

Le Roux, X., Barbault, R., Baudry, J., Burel, F., Doussan, I., Garnier, E., Herzog, F., Lavorel, S., Lifran, R., Roger-Estrade, J., Sarthou, J. P., and Trommetter, M., (Eds.) (2008). Agriculture et biodiversité. Valoriser les synergies. Expertise scientifique collective, synthèse du rapport. Paris, INRA .France 112 p.

Mac Neely, J.A. *et al*. (2005). Ecosystems and Human Well-Being, vol 3, Policy responses, Chopra, K., Leemans, R., Kumar, P., Simons, H., Eds (Millenium assessment, island press, Washington DC), 119-172.

Mitchell, G, May, A., and McDonald, A. (1995). PICABUE: a methodological framework for the development of indicators of sustainable development. *International Journal od Sustainable Development and worls Ecology*, 2:104-123.

Navarro, M., Rochs, F., Prussia, S.E., Kanlayanarat, S., and McGlasson, B. (2008). Using soft-systems methodology to plan advanced academic programs in post-harvest technology in Thailand. 24TH annual conference of the assocation for international agricultural and extension education. Guacimo (costa-rica), 10-15 mars 2008.

Patton, M. Q. (ed) (1975). *Alternative evaluation research paradigms*. Grand Foks: University of North Dakota Press.

Patton, M. Q. (ed) (2008). *Utilization focused evaluation*, California, 4th Edition, Sage.

Pawson, R. andTilley, N. (eds) (1997). *Realistic Evaluation*, London: Sage.

Pawson, R. (2006). Evidence-based policy : a realistic perspective. London:Sage publications.

Primdahl J., Peco B., Schramek J., Anderse, E., Onate JJ., 2003, Environmental effects of agri-environmental schemes in Western Europe, Journal of Environmental Management 67 (2003) 129–138

Rochs, F., and Navarro, M. (2008). Soft System Methodology: an intervention strategy. *Journal of International Agricultural and extension education*, 15(3): 95-99.

Rossi, P.H. and Freeman, H.E. (eds) (1993). *H.E. Evaluation : a systématic approach*. 5th Edition. Newbury Park, CA : Sage.

Salner, M., (2000). Beyond Checkland & Scholes: Improving SSM, *Occasional Papers on Systemic Development*, , 11: 23-44.

Scriven, M. (1998). Minimalist theory : the least theory that practice requires. American *Journal of Evaluation*, 19(11): 57-70.

Shadish, W.R., Cook, T.D., and Leviton, L.C. (eds) (1991). *Foundations of Program Evaluation Theories of Practice.* Newbury Park: Sage publications.

Van der Suijs, J., Douguet, J.-M., O'Connor, M., Guimaraes Periera, A., Quintana, S.C., Maxim, L., and Ravetz, J. (2008). Qualité de la connaissance dans un processus délibératif. *Nature, Science, Société*, 16: 265-73.

Weber, M. (1919 / 1959). Le savant et le politique. Book resulting from two conférences prononced in 1919 Wissenschaft als Beruf and Politik als Beruf. 10/18, 220p

Van der Werf, H., and Petit, J. (2002). Evaluation of the environmental impact of agriculture at the farm level: a comparison and analysis of 12 indicators-based method, *Agriculture, Ecosystems and Environment*, 93: 131-145.